February 3, 2020. **Molecular Data I:** General introduction; types of molecular data (DNA hybridization; allozymes; restriction sites, DNA sequences, ESTs; comparative genomics)

Required reading: Maddison, W. P. and D.R. Maddison. 2019. Mesquite: a modular system for evolutionary analysis. Version 3.61, Chapter on Analyzing Molecular Data: https://www.mesquiteproject.org/Analyzing%20Molecular%20Data.html?TaxaTreesCharacters

I. Techniques - kinds of data
- -- intrinsically distance-based data:
  - -- immunology (cross reaction of antibodies)
  - -- DNA - DNA hybridization
  - -- AFLPs - RAPDs - DNA fingerprinting - RADSeq
  - -- microsatellites
- -- character-based data:
  - -- allozymes
  - -- restriction enzyme sites
  - -- sequencing methods
    - -- direct
    - -- cloning
    - -- PCR/ next gen approaches
  - -- genomic data (gene arrangement)

Properties of a good marker, as compared between molecules (i.e., DNA sequence data) and morphology.

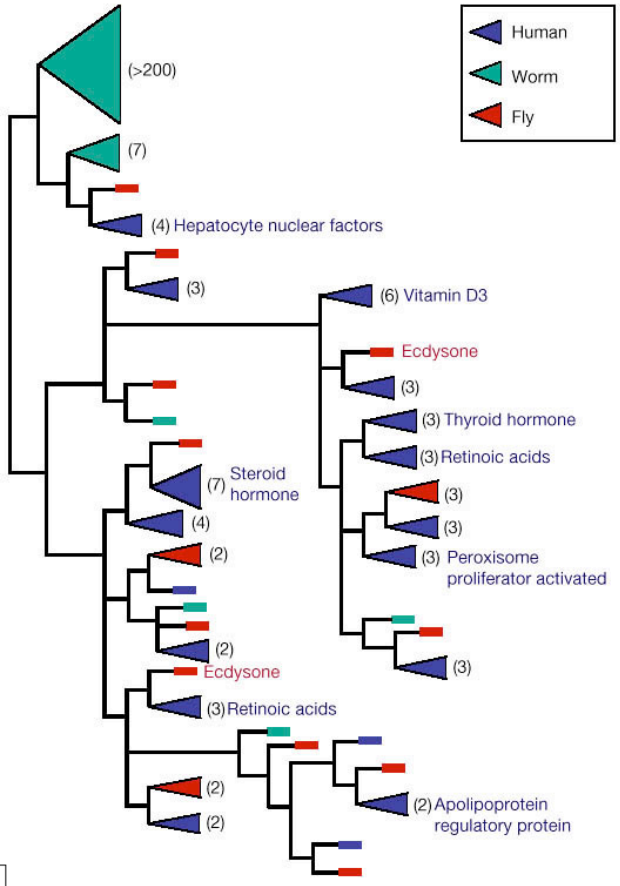
	molecules	morphology
1) COMPLEXITY AND COMPARABILITY	–	+
2) DISCRETE STATES	+	–
3) HERITABILITY	+	–
4) INDEPENDENCE	?	?
5) LOW RATE OF CHANGE (λ)	?	?
6) MANY POSSIBLE CHARACTER STATES	–	+

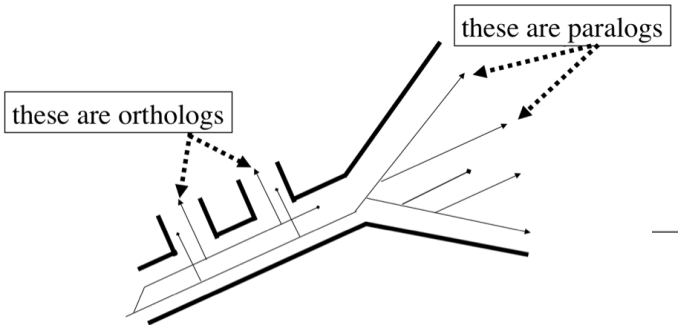
II. Special features of molecular data
- -- purported advantages:
  - -- closer to (or equal to) the genetic information.
  - -- huge numbers of potential characters, especially useful in organisms with simple morphology.
  - -- ability to homologize across very broad groups.
  - -- independence from morphological characters which are perhaps more subject to adaptive convergence.
  - -- ability to model or weight, because of relatively simple models of change.
  - -- $$$.
- -- purported weaknesses:
  - -- simplicity of characters (i.e., no ontogeny, few possible character states) leading to special problems with homoplasy.
  - -- sampling problems.
  - -- fossil taxa generally can't be included.
  - -- highly conserved regions, used to reconstruct deep branching points, are perhaps *more* subject to adaptive convergence.
  - -- $$$.

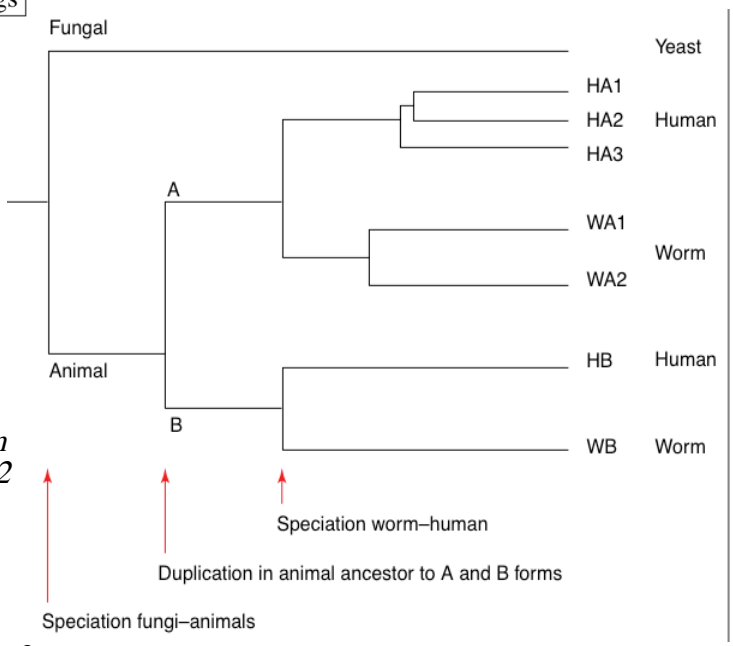
III. Issues arising with molecular data:
-- homology (including alignment problems - covered next lecture)
-- what is a character?
-- nucleotide positions
-- character correlations
-- structural rearrangements (i.e., deletions, inversions) - more below
-- allozymes
-- restriction sites:
--RFLP's
--mapping
-- microsatellites
-- SNPs

-- weighting/modeling issues:
-- gains versus losses
-- transitions versus transversions
-- purines A G
-- pyrimidines C T U
-- codon position bias
-- compensatory substitutions in RNA (due to secondary structure)

IV. Comparing genomes
-- synteny, rearrangements, insertions/deletions
-- exon shuffling
-- the gene "annotation" problem
-- multigene families
-- paralogy vs orthology
-- the fate of duplicated genes: ghost genes, subfunctionalization







"inparalogs" vs. "outparalogs"
Erik Sonnhammer. *Trends in Genetics Vol.18 No.12, 2002*

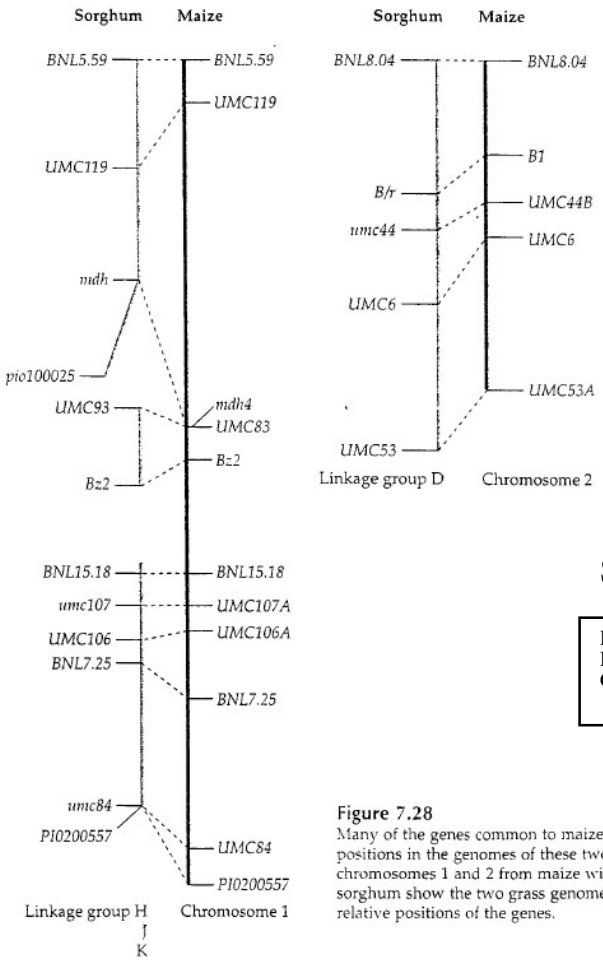
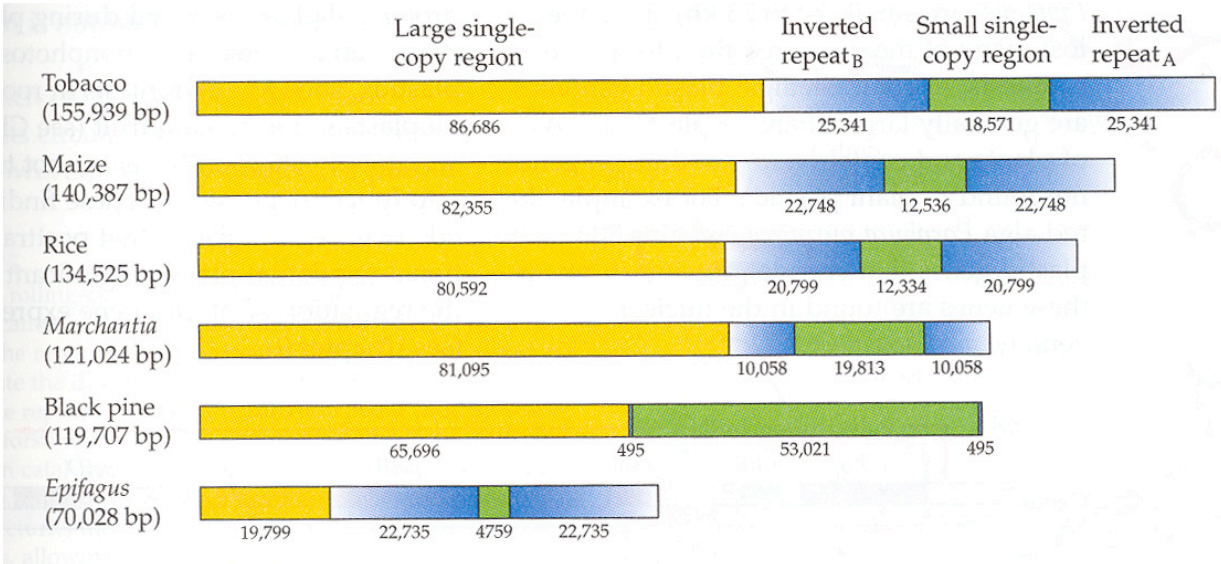


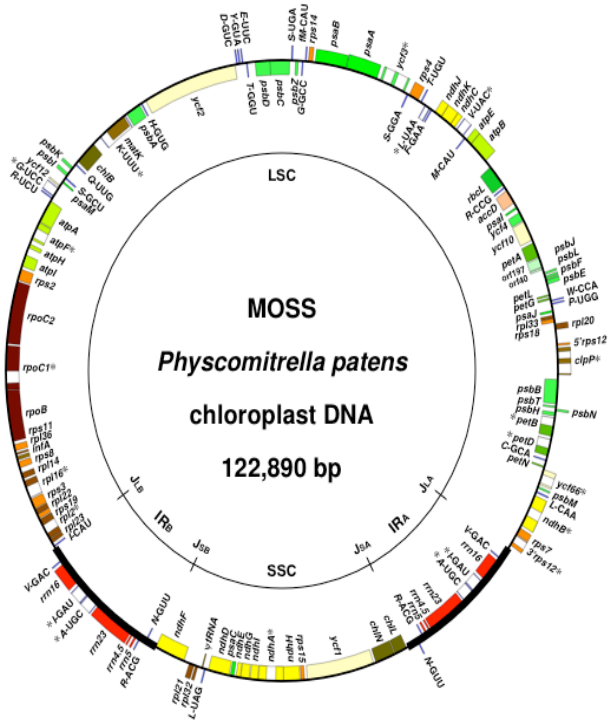
Figure 7.28
Many of the genes common to maize and sorghum are organized into similar positions in the genomes of these two plants. A comparison of genetic maps of chromosomes 1 and 2 from maize with the maps of several linkage groups from sorghum show the two grass genomes similarly organized with respect to the relative positions of the genes.

Synteny

From: Biochemistry & Molecular Biology of Plants. edited by Bob B. Buchanan, Wilhelm Gruissem, Russell L. Jones. 2000.



Multiple gene inversion characters across Green Plants:



5 gene inversion in LSC
Pteridophytes s.s.

12 gene inversion and 13 gene inversion and translocation in LSC
Chaetosphaeridium

33 gene inversion in LSC
Oenothera

75 gene inversion in LSC
Physcomitrella

D.G. Kelch, A. Driskell, and B.D. Mishler. 2004. Inferring phylogeny using genomic characters: a case study using land plant plastomes, *In* B. Goffinet, V. Hollowell, and R. Magill (eds.), Molecular Systematics of Bryophytes [Monographs in Systematic Botany 98], pp. 3-12. Missouri Botanical Garden Press, St. Louis.

V. Recommendations (Mishler's Aphorisms):

-- treat these data as any other; if the object is phylogeny reconstruction, use phylogenetic methods.

-- include all available data in an analysis, even if your own focus has been on molecules; it makes no sense to ignore older data just because newer data have been generated.

-- be wary of consensus tree approaches; they may be worthwhile as part of the analysis, but it is probably best to combine all putative homologies into one matrix (perhaps with weighting if this can be independently justified).

-- for reconstructing deep splits, it is much better to sequence portions of several different genes, scattered around the nuclear and organelle genomes, than it is to concentrate on extensive sequencing of a single gene (because of the problem of tight selective constraints on any one highly conserved region). Or for that matter, use morphology or genome structure.

-- it is probably better to break large surveys down into reasonable local analyses, to avoid spurious homoplasy (e.g., instead of putting all eukaryotes into one huge matrix, work on relationships within smaller, a priori justified monophyletic groups, and later link those groups together using archetypes: "compartmentalization" (Mishler, 2005).

-- molecular evolutionary studies and phylogeny reconstruction using molecules are two very different goals; for the former purpose, one should use phylogenies based on morphology (and other characters, perhaps including molecules -- but not the molecules that are being studied evolutionary).